Imitation Learning of Robot Policies by Combining Language, Vision and Demonstration

Simon Stepputtis[1]**, Joseph Campbell**[1]**, Mariano Phielipp**[2]**, Chitta Baral**[1]**, Heni Ben Amor**[1]
[1] School of Computing, Informatics, and Decision Systems Engineering, ASU
`{sstepput, jacampb1, chitta, hbenamor}@asu.edu`
[2] Intel AI Lab
`mariano.j.phielipp@intel.com`

Abstract

In this work we propose a novel end-to-end imitation learning approach which combines natural language, vision, and motion information to produce an abstract representation of a task, which in turn is used to synthesize specific motion controllers at run-time. This multimodal approach enables generalization to a wide variety of environmental conditions and allows an end-user to direct a robot policy through verbal communication. We empirically validate our approach with an extensive set of simulations and show that it achieves a high task success rate over a variety of conditions while remaining amenable to probabilistic interpretability.

1   Introduction

A significant challenge when designing robots to operate in the real world lies in the generation of control policies that can adapt to changing environments. Programming such policies is a labor and time-consuming process which requires substantial technical expertise. Imitation learning [Schaal, 1999], is an appealing methodology that aims at overcoming this challenge – instead of complex programming, the user only provides a set of demonstrations of the intended behavior. These demonstrations are consequently distilled into a robot control policy by learning appropriate parameter settings of the controller. Popular approaches to imitation, such as Dynamic Motor Primitives (DMPs) [Ijspeert et al., 2013] or Gaussian Mixture Regression (GMR) [Calinon, 2009] largely focus on motion as the sole input and output modality, i.e., joint angles, forces or positions. Critical semantic and visual information regarding the task, such as the appearance of the target object or the type of task performed, is not taken into account during training and reproduction. The result is often a limited generalization capability which largely revolves around adaptation to changes in the object position. While imitation learning has been successfully applied to a wide range of tasks including table-tennis Mülling et al. [2013], locomotion Chalodhorn et al. [2007], and human-robot interaction Amor et al. [2014] an important question is how to incorporate language and vision into a differentiable end-to-end system for complex robot control.

In this paper, we present an imitation learning approach that combines language, vision, and motion in order to synthesize natural language-conditioned control policies that have strong generalization capabilities while also capturing the semantics of the task. We argue that such a multi-modal teaching approach enables robots to acquire complex policies that generalize to a wide variety of environmental conditions based on descriptions of the intended task. In turn, the network produces control parameters for a lower-level control policy that can be run on a robot to synthesize the corresponding motion. The hierarchical nature of our approach, i.e., a high-level policy generating the parameters of a lower-level policy, allows for generalization of the trained task to a variety of spatial, visual and contextual changes.

Problem Statement: In order to outline our problem statement, we contrast our approach to Imitation learning [Schaal, 1999] which considers the problem of learning a policy π from a given set of demonstrations $\mathcal{D} = \{\mathbf{d}^0, .., \mathbf{d}^m\}$. Each demonstration spans a time horizon T and contains information about the robot's states and actions, e.g., demonstrated sensor values and control inputs at each time step. Robot states at each time step within a demonstration are denoted by $\mathbf{x}_t$. In contrast to other imitation learning approaches, we assume that we have access to the raw camera images of the robot $\boldsymbol{I}_t$ at teach time step, as well as access to a verbal description of the task in natural language. This description may provide critical information about the context, goals or objects involved in the task and is denoted as $\mathbf{s}$. Given this information, our overall objective is to learn a policy π which imitates the demonstrated behavior, while also capturing semantics and important visual features. After training, we can provide the policy $\pi(\mathbf{s}, \boldsymbol{I})$ with a different, new state of the robot and a new verbal description (instruction) as parameters. The policy will then generate the control signals needed to perform the task which takes the new visual input and semantic context int o account.

2 Background

A fundamental challenge in imitation learning is the extraction of policies that do not only cover the trained scenarios, but also generalize to a wide range of other situations. A large body of literature has addressed the problem of learning robot motor skills by imitation [Argall et al., 2009], learning functional [Ijspeert et al., 2013] or probabilistic [Maeda et al., 2014] representations. However, in most of these approaches, the state vector has to be carefully designed in order to ensure that all necessary information for adaptation is available. Neural approaches to imitation learning [Pomerleau, 1989] circumvent this problem by learning suitable feature representations from rich data sources for each task or for a sequence of tasks [Burke et al., 2019, Hristov et al., 2019, Misra et al., 2018]. Many of these approaches assume that either a sufficiently large set of motion primitives is already available or that a taxonomy of the task is available, i.e., semantics and motions are not trained in conjunction. The importance of maintaining this connection has been shown in Chang et al. [2019], allowing the robot to adapt to untrained variations of the same task. To learn entirely new tasks, meta-learning aims at learning policy parameters that can quickly be fine-tuned to new tasks [Finn et al., 2017]. While very successful in dealing with visual and spatial information, these approaches do not incorporate any semantic or linguistic component into the learning process. Language has shown to successfully generate task descriptions in Arumugam et al. [2019] and several works have investigated the idea of combining natural language and imitation learning: Nicolescu and Mataric [2003], Gemignani et al. [2015], Cederborg and Oudeyer [2013], Mericli et al. [2014], Sugita and Tani [2005]. However, most approaches do not utilize the inherent connection between semantic task descriptions and low-level motions to train a model.

Our work is most closely related to the framework introduced in Tellex et al. [2014], which also focuses on the symbol grounding problem. More specifically, the work in Tellex et al. [2014] aims at mapping perceptual features in the external world to constituents in an expert-provided natural language instruction. Our work approaches the problem of generating dynamic robot policies by fundamentally combining language, vision, and motion control in to a single differentiable neural network that can learn the cross-modal relationships found in the data with minimal human feature engineering. Unlike previous work, our proposed model is capable of directly generating complex low-level control policies from language and vision that allow the robot to reassemble motions shown during training.

3 Multimodal Policy Generation via Imitation

We motivate our approach with a simple example: consider a binning task in which a 6 DOF robot has to drop an object into one of several differently shaped and colored bowls on a table. To teach this task, the human demonstrator does not only provide a kinesthetic demonstration of the desired trajectory, but also a verbal command, e.g., *"Move towards the blue bowl"* to the robot. In this example, the trajectory generation would have to be conditioned on the *blue* bowl's position which, however, has to be extracted from visual sensing. Our approach automatically detects and extracts these relationships between vision, language, and motion modalities in order to make best usage of contextual information for better generalization and disambiguation.

Figure 1 (left) provides an overview of our method. Our goal is to train a deep neural network that can take as input a task description $\mathbf{s}$ and and image $\boldsymbol{I}$ and consequently generates robot controls. In
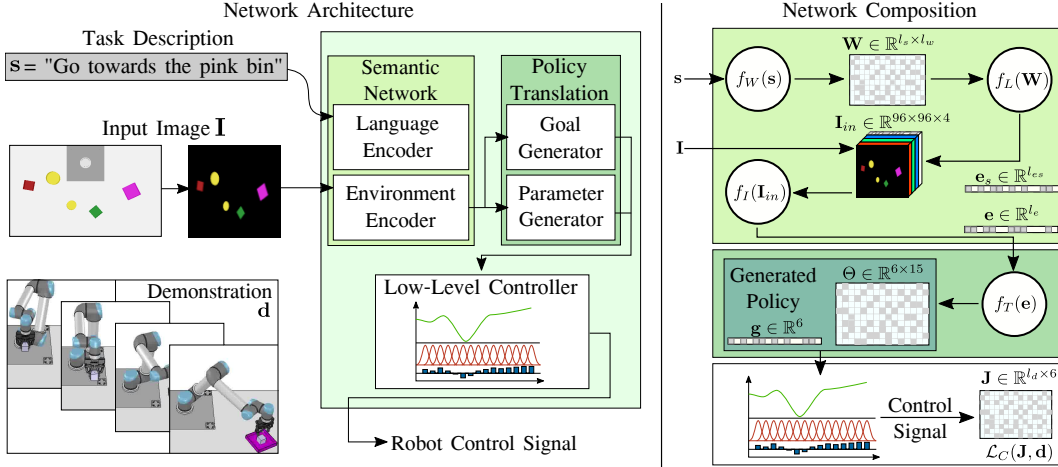
Figure 1: Network architecture overview. The network consists of two parts, a high-level semantic network and a low-level control network. Both networks are working seamlessly together and are utilized in an End-to-End fashion.

the remainder of this paper, we will refer to our network as the Multimodal Policy Network (MPN). Rather than immediately producing control signals, the MPN will generate the parameters for a lower-level controller. This distinction allows us to build upon well-established control schemes in robotics and optimal control. In our specific case, we use the widely used Dynamic Motor Primitives [Ijspeert et al., 2013] as a lower-level controller for control signal generation.

In essence, our network can be divided into three parts. The first part, the semantic network, is used to create a task embedding e from the input sentence s and environment image I. In a first step, the sentence s is tokenized and converted into a sentence matrix $W \in \mathbb{R}^{l_s \times l_w} = f_W(s)$ by utilizing pre-trained Glove word embeddings [Pennington et al., 2014] where l_s is the padded-fixed-size length of the sentence and l_w is the size of the glove word vectors. To extract the relationships between the words, we use use multiple CNNs $e_s = f_L(W)$ with filter size $n \times l_w$ for varying n, representing different n-gram sizes [Yang et al., 2015]. The final representation is built by flattening the individual n-grams with max-pooling of size $(l_s - n_i + 1) \times l_w$ and concatenating the results before using a single perceptron to detect relationships between different n-grams. In order to combine the sentence embedding e_s with the image, it is concatenated as a fourth channel to the input image I. The task embedding e is produced with three blocks of convolutional layers, composed of two regular convolutions, followed by a residual convolution [He et al., 2015] each.

In the second part, the policy translation network is used to generate the task parameters $\Theta \in \mathcal{R}^{o \times b}$ and $g \in \mathcal{R}^o$ given a task embedding e where o is the number of output dimensions and b the number of basis functions in the DMP:

$$\Theta, g = f_T(e) = f_G\left(\text{ReLU}\left(W_G e + b_G\right)\right), f_H\left(\text{ReLU}\left(W_G e + b_G\right)\right) \tag{1}$$

where $f_G()$ and $f_H()$ are multilayer-perceptrons that use e after being processed in a single perceptron with weight W_G and bias b_G. These parameters are then used in the third part of the network, which is a DMP [Schaal, 1999], allowing us leverage a large body of research regarding their behavior and stability, while also allowing other extensions of DMPs [Amor et al., 2014, Paraschos et al., 2013, Khansari-Zadeh and Billard, 2011] to be incorporated to our framework.

4 Results

We evaluate our model in a simulated binning task in which the robot is tasked to place a cube into a bowl as outlined by the verbal command. Each environment contains between three and five objects differentiated by their size (small, large), shape (round, square) and color (red, green, blue, yellow, pink), totalling in 20 different objects. Depending on the generated scenario, combinations of these three features are necessary to distinguish the targets from each other, allowing for tasks of varying complexity.
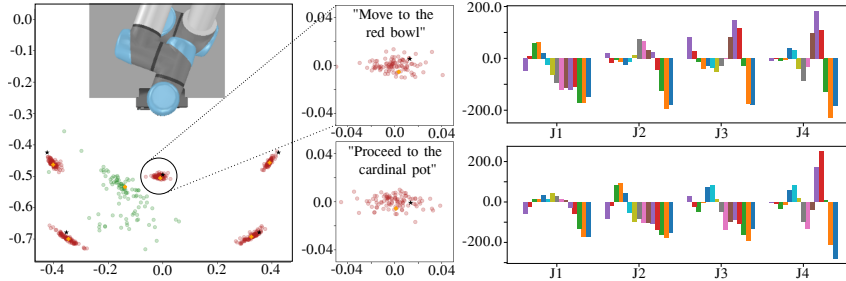
3

Figure 2: Results for placing an object into bowls at different locations: (Left) Stochastic forward passes allow the model to estimate its certainty about the validity of a task. (Right) Generated weights Θ for four joints of the DMP shown for two objects close and far away of the robot.

To train our model, we generated a dataset of 20,000 demonstrated 7 DOF trajectories (6 robot joints and 1 gripper dimension) in our simulated environment together with a sentence generator capable of creating natural task descriptions for each scenario. In order to create the language generator, we conducted an human-subject study to collect sentence templates of a placement task as well as common words and synonyms for each of the used features. By utilising these data, we are able to generate over 180,000 unique sentences, depending on the generated scenario.

The generated parameters of the low-level DMP controller – the weights and goal position – must be sufficiently accurate in order to successfully deliver the object to the specified bin. On the right side of Figure 2, the generated weights for the DMP are shown for two tasks in which the target is close and far away from the robot, located at different sides of the table, indicating the robots ability to generate differently shaped trajectories. The accuracy of the goal position can be seen in Figure 2(left) which shows another aspect of our approach: By using stochastic forward passes [Gal and Ghahramani, 2015] the model can return an estimate for the validity of a requested task in addition to the predicted goal configuration. The figure shows that the goal position of a red bowl has a relatively small distribution independently of the used sentence or location on the table, where as an invalid target (green) produces a significantly larger distribution, indicating that the requested task may be invalid.

To test our model, we generated 500 new scenario testing each of the three features to identify the correct target among other bowls. A task is considered to be successfully completed when the cube is withing the boundaries of the targeted bowl. Bowls have a bounding box of 12.5 and 17.5cm edge length for the small and large variant, respectively. Our experiments showed that using the objects color or shape to uniquely identify an object allows the robot successfully complete the binning task in 97.6% and 96.0% of the cases. However, using the shape alone as a unique identifier, the task could only be completed in 79.0% of the cases. We suspect that the loss of accuracy is due to the low image resolution of the input image, preventing the network from reliably distinguishing the object shapes. In general, our approach is able to actuate the robot with an target error well below 5cm, given the target was correctly identified.

5   Conclusion and Future Work

In this work, we presented an imitation learning approach combining language, vision, and motion. A neural network architecture called Multimodal Policy Network was introduced which is able to learn the cross-modal relationships in the training data and achieve high generalization and disambiguation performance as a result. Our experiments showed that the model is able to generalize towards different locations and sentences while maintaining a high success rate of delivering an object to a desired bowl. In addition, we discussed an extensions of the method that allow us to obtain uncertainty information from the model by utilizing stochastic network outputs to get a distribution over the belief.

The modularity of our architecture allows us to easily exchange parts of the network. This can be utilized for transfer learning between different tasks in the semantic network or transfer between different robots by transferring the policy translation network to different robots in simulation, or to bridge the gap between simulation and reality.

4

References

Heni Ben Amor, Gerhard Neumann, Sanket Kamthe, Oliver Kroemer, and Jan Peters. Interaction primitives for human-robot cooperation tasks. In *2014 IEEE international conference on robotics and automation (ICRA)*, pages 2831–2837. IEEE, 2014.

Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

Dilip Arumugam, Siddharth Karamcheti, Nakul Gopalan, · Edward C Williams, Mina Rhee, · Lawson, L S Wong, and Stefanie Tellex. Grounding natural language instructions to semantic goal representations for abstraction and generalization. *Autonomous Robots*, 43:449–468, 2019. doi: 10.1007/s10514-018-9792-8. URL https://doi.org/10.1007/s10514-018-9792-8.

Michael Burke, Svetlin Penkov, and Subramanian Ramamoorthy. From explanation to synthesis: Compositional program induction for learning from demonstration. feb 2019. doi: 10.15607/RSS. 2019.XV.015. URL http://arxiv.org/abs/1902.10657http://dx.doi.org/10.15607/RSS.2019.XV.015.

Sylvain Calinon. *Robot programming by demonstration*. EPFL Press, 2009.

Thomas Cederborg and Pierre-Yves Oudeyer. From language to motor gavagai: Unified imitation learning of multiple linguistic and nonlinguistic sensorimotor skills. *IEEE Trans. on Auton. Ment. Dev.*, 5(3):222–239, September 2013. ISSN 1943-0604. doi: 10.1109/TAMD.2013.2279277. URL http://dx.doi.org/10.1109/TAMD.2013.2279277.

Rawichote Chalodhorn, David B Grimes, Keith Grochow, and Rajesh PN Rao. Learning to walk through imitation. In *IJCAI*, volume 7, pages 2084–2090, 2007.

Jonathan Chang, Nishanth Kumar, Sean Hastings, Aaron Gokaslan, Diego Romeres, Devesh K Jha, Daniel Nikovski, George Konidaris, and Stefanie Tellex. Learning Deep Parameterized Skills from Demonstration for Re-targetable Visuomotor Control. Technical report, 2019.

Chelsea Finn, Tianhe Yu, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. One-shot visual imitation learning via meta-learning. In Sergey Levine, Vincent Vanhoucke, and Ken Goldberg, editors, *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pages 357–368. PMLR, 13–15 Nov 2017. URL http://proceedings.mlr.press/v78/finn17a.html.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2015.

Guglielmo Gemignani, Emanuele Bastianelli, and Daniele Nardi. Teaching robots parametrized executable plans through spoken interaction. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '15, pages 851–859, Richland, SC, 2015. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-1-4503-3413-6. URL http://dl.acm.org/citation.cfm?id=2772879.2773262.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Yordan Hristov, Daniel Angelov, Michael Burke, Alex Lascarides, and Subramanian Ramamoorthy. Disentangled Relational Representations for Explaining and Learning from Demonstration. jul 2019. URL http://arxiv.org/abs/1907.13627.

Auke Jan Ijspeert, Jun Nakanishi, Heiko Hoffmann, Peter Pastor, and Stefan Schaal. Dynamical movement primitives: learning attractor models for motor behaviors. *Neural computation*, 25(2): 328–373, 2013.

S Mohammad Khansari-Zadeh and Aude Billard. Learning stable nonlinear dynamical systems with gaussian mixture models. *IEEE Transactions on Robotics*, 27(5):943–957, 2011.

Guilherme Maeda, Marco Ewerton, Rudolf Lioutikov, Heni Ben Amor, Jan Peters, and Gerhard Neumann. Learning interaction for collaborative tasks with probabilistic movement primitives. In *Humanoid Robots (Humanoids), 2014 14th IEEE-RAS International Conference on*, pages 527–534. IEEE, 2014.

Cetin Mericli, Steven D. Klee, Jack Paparian, and Manuela Veloso. An interactive approach for situated task specification through verbal instructions. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '14, pages 1069–1076, Richland, SC, 2014. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-1-4503-2738-1. URL `http://dl.acm.org/citation.cfm?id=2617388.2617416`.

Dipendra Misra, John Langford, and Yoav Artzi. Mapping Instructions and Visual Observations to Actions with Reinforcement Learning, jan 2018. URL `https://arxiv.org/abs/1704.08795`.

Katharina Mülling, Jens Kober, Oliver Kroemer, and Jan Peters. Learning to select and generalize striking movements in robot table tennis. *The International Journal of Robotics Research*, 32(3): 263–279, 2013.

Monica N. Nicolescu and Maja J. Mataric. Natural methods for robot task learning: Instructive demonstrations, generalization and practice. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*, AAMAS '03, pages 241–248, New York, NY, USA, 2003. ACM. ISBN 1-58113-683-8. doi: 10.1145/860575.860614. URL `http://doi.acm.org/10.1145/860575.860614`.

Alexandros Paraschos, Christian Daniel, Jan R Peters, and Gerhard Neumann. Probabilistic movement primitives. In *Advances in neural information processing systems*, pages 2616–2624, 2013.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL `http://www.aclweb.org/anthology/D14-1162`.

Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In *Advances in neural information processing systems*, pages 305–313, 1989.

Stefan Schaal. Is imitation learning the route to humanoid robots? *Trends in cognitive sciences*, 3(6): 233–242, 1999.

Yuuya Sugita and Jun Tani. Learning Semantic Combinatoriality from the Interaction between Linguistic and Behavioral Processes. Technical report, 2005.

Stefanie Tellex, Pratiksha Thaker, Joshua Joseph, and Nicholas Roy. Learning perceptually grounded word meanings from unaligned parallel data. *Machine Learning*, 94(2):151–167, Feb 2014. ISSN 1573-0565. doi: 10.1007/s10994-013-5383-2. URL `https://doi.org/10.1007/s10994-013-5383-2`.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. Stacked attention networks for image question answering. *CoRR*, abs/1511.02274, 2015. URL `http://arxiv.org/abs/1511.02274`.